Processing Obtained Email Data By Using Naïve Bayes Learning Algorithm

Valery V. Pelenko, Aleksandr V. Baranenko

St. Petersburg Research University of Information Technologies, Mechanics and Optics.
Institute of Refrigeration and Biotechnology
9 Lomonosova St., St Petersburg, 191002, Russia

ABSTRACT

This paper gives a basic idea how various machine learning techniques may be applied towards processing the data from DEA services to find out whether people use these services for legitimate or non-legitimate purposes.

KEYWORDS

Spam, ham, emails, communications, machine learning, Naïve Bayes, DEA services.

1. INTRODUCTION

In this paper we are characterizing a set of found email data using a method of supervised learning – Naïve Bayes. There are lots of DEA services, but we decided to get the data from most commonly used DEA services [1]. The data set was chosen from five independent disposable email address (DEA) services: dispostable, mailinator, mytrashmail, staging and tempemail. The DEA allow users to receive the emails without creating an account. Not much is known how people use these services; in fact, there are so many legitimate and non-legitimate purposes that people use these services for. Since the accounts don't require the password to access them, this data becomes public. Some users may not realize how public this data are and that anyone else can purposely or accidentally access this private information. We decided to go ahead and categorize the found email data and split it up in four main categories. Our goal is to find out how much of the information that is stored in the DEA services is legitimate. At the same time, the main purpose for this paper was to break down the data usage and present the statistics of the DEA services.

We found that about 72% of the data was tagged 'spam' comparing to only 6.7% of legitimate data which was tagged 'ham'. The remaining part was split unevenly between 6.2% of 'other' and 15.1% of 'Non English'. We used Naïve Bayes to split the data into the categories mentioned above.

During thus project, we were dealing with the data that was obtained from five various independent DEA services. DEA services are such services that do not require a user to create an account in order to receive an email to any email address supported by the service. Thus, all the data becomes public due to no identification required. Since the DEA services are used by many users from all over the world for a variety of purposes, then the data that was obtained such services is obviously to be very diverse. It is very important to understand that users may use DEA services for both legitimate and non-legitimate purposes [2]. Since all the data is public and

DOI : 10.5121/ijcsit.2014.6111

some users use the DEA services for legitimate purposes, these users may not realize how public the data are and what risk they take. This is the main factor why DEA services may not be very helpful as they seem to be at first. In order to protect personal information and avoid an identity theft, the users should be informed beforehand what risk they take. Some non-legitimate users may find this personal data such as: name, address, SSN, credit card number, and others very useful and the victim can suffer adverse consequences if they are held accountable for the perpetrator's actions.

The main purpose of the paper was to characterize this found email data. Once we are done splitting the data into different categorize, we can look at the informative data. The informative data would be categorized as 'ham'. A closer look at the messages in the 'ham' category will benefit us in some ways such are presented below:

- Understand what people use DEA services for legitimate purposes.
- What did force these users use DEA service over the standard email account?
- Is there any risk that people take when using DEA services for legitimate purposes, if there is, how high is that risk?

After processing all the data, We will be able to demonstrate how much of this data are used for legitimate purposes, non-legitimate purposes and any other purposes (if any); show whether the users take any risk of loss of any personal information when using DEA services

2. FOUND EMAIL DATA SETS

Disposable email addressing (DEA) is a way of sharing and managing email addressing [3]. DEA allows user to set up a new, unique email address for every contact or use an already existing email address that only requires having a username to access the existing account in order to make a connection between the sender and the recipient.

Our main idea is that the DEA services are mostly used by people who try to avoid using their personal email accounts due to various factors including receiving spam [3,4]. Thus, these people prefer to use DEA services to stay anonymous.

The main advantage of using the DEA services is that there is no need to register by using your real credentials and, of course, these services are free of charge. The user is given a choice to select any name for the email address and use it, even if it was used before [5].

As any other service, there are also disadvantages of using the DEA services. If this account has been used before, then the user will be able to track all the messages in this account and some private information may become public. Many forums and legitimate services filter out messages sent from DEA domains.

3. METHODOLOGY

The primary work for this paper was to characterize the found email data from various email services listed below:

- Dispostable
- Mailinator
- Mytrashmail
- Staging
- Tempemail

We have competed two separate stages of characterizing the data. We were mainly interested in breaking down the category of "ham" email versus "spam" category. The categories used at the first and second stages are listed below:

Stage 1:

- Spam

  The messages considered to be under the category of spam if it's obvious that the recipient has no pre-existing relations with the sender. For example, any message that looks like an advertisement and it was sent to thousands of people simultaneously, this message will be considered as spam [4].

- Ham

  The messages considered to be under the category of ham if the user seems to have pre-existing relation such as the message seems to indicate a specific action taken by the person to cause it (an invoice for purchased products, or a response mentioning that a purchase could not go through etc.) will be considered ham [4].

- Non English

  The messages considered to be non English if the original language in what the message was written is other than English, e.g., Russian, Spanish, Italian, Chinese, Arabic, Hebrew, Ukrainian (the listed languages were found in the found email data set).

- Other

  The messages considered to be under the category of other if the messages couldn't be assigned to any other category mentioned above. For example, an email with no text in it and a plain black background would be described as other.

- Errors

  The messages considered to be under the category of other if the messages couldn't be parsed by program or broken files.

Stage 2:

- Buying

  The messages considered to be under the category of buying if the email that was received by the recipient mentions that he or she has purchased something. For example, an email received from paypal service would be considered as buying email.

- Signing Up

  The messages considered to be under the category of signing up if the email that was received by the percipient shows that the user has subscribed to some service. For example, an email received from the forum web sites is considered as signing up email.

- Non Ham (Spam)

  The definition provided above in stage 1.

- Personal

  The messages considered to be under the category of personal if the user has requested a subscription from the social websites. For example, an email received from Linkedin is considered as a personal email.

The found email data, as mention above, was taken from five independent DEA services. Please refer to the table to see what the size of the data was from each service.

DEA Service	**Size of the data (GB)**	**% of data**
Dispostable	2.1	24.21%
Mailinator	0.009	0.010%
Mytrashmail	0.22	0.253%
Staging	0.095	1.095%
Tempemail	6.25	74.432%
Overall	**8.674**	**100%**

There are two main sets of machine learning techniques that can be used for classifying the data: supervised and unsupervised learning [6]. The supervised learning is when the data is tagged before the algorithm makes any further decisions. On the other hand, if there is no input to the algorithm, then unsupervised learning has to be used. Algorithms in the group of unsupervised learning find the similarities and/or correlations in the data and require no input to classify the data.

There was a choice of using either supervised or unsupervised learning. Since the data set of found email data was 8.67 GB, WE decided to use supervised learning. Naïve Bayes method was chosen to be used as a method for supervised learning [6].

For the purposes of the describing the data, we decided to use the Naïve Bayes algorithm which is suitable for handling big sets of data, in our example the data set is 8.764 GB and diverse data.

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with naive independence assumptions. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, Naïve Bayes classifiers can be trained very efficiently in a supervised learning setting [6]. In many practical applications, parameter estimation for Naïve Bayes models uses the method of maximum likelihood; in other words, one can work with the Naïve Bayes model without believing in Bayesian probability or using any Bayesian methods [7].

Abstractly, the probability model for a classifier is a conditional model: $P(C \mid F_1,...,F_n)$ over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F_1 through F_n. The problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible [6]. We therefore reformulate the model to make it more tractable.

Using Bayes' theorem, this can be written as following:

$$P(C \mid F_1,...,F_n) = \frac{p(C)p(F_1,...,F_n \mid C)}{p(F_1,...,F_n)}$$

For the purpose of understanding, in simple English the equation provided above can be written as following:

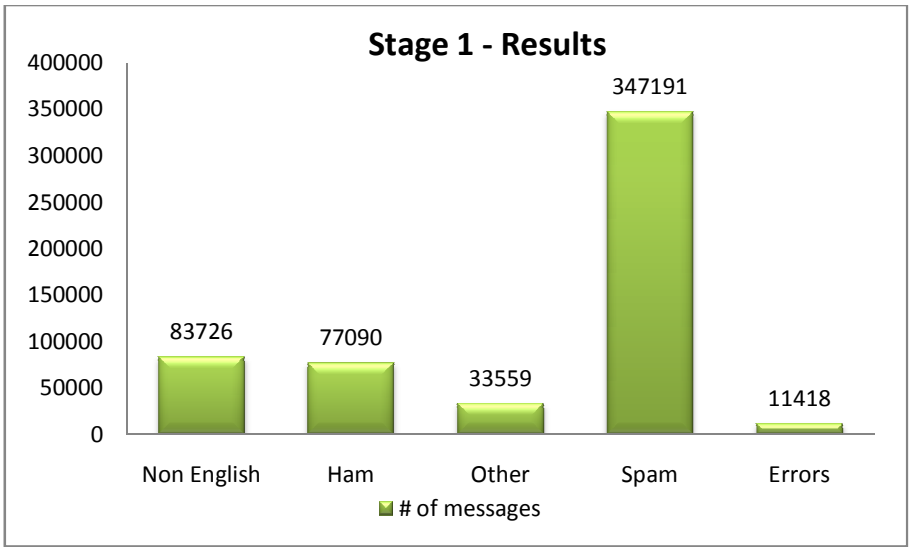
$$posterior = \frac{prior * likelihood}{evidence}$$

4. RESULTS

Due to two factors such as diversity of the data and the size of the data, the algorithm was run twice in order to split the data. At the first stage the following tags were introduced into the model. The description of each tag is provided in the introduction section. The size of data considered at the first stage was 8.76 GB.

1. Spam
2. Ham
3. Non English
4. Other
5. Errors

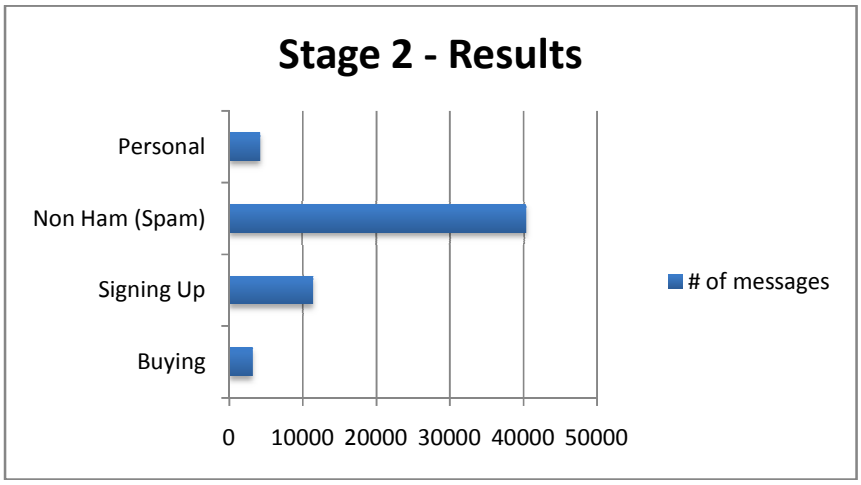
After training the algorithm, the size of the database which contains the words and its probabilities was 1878 KB (234 messages). The results obtained after the first stage are as following:



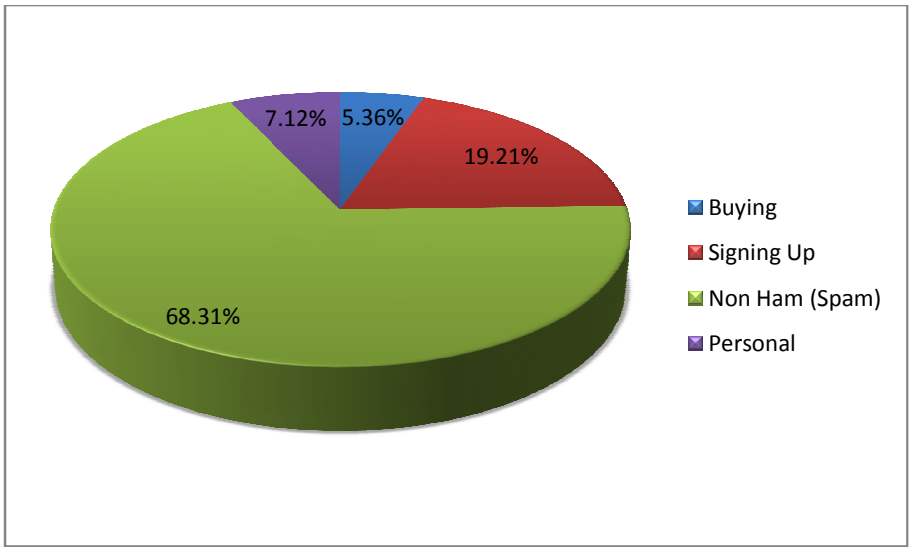
Since the main idea was to break down the category of ham email, the algorithm at the second stage filtered out the emails that were tagged as 'Ham' after completing the first stage. In order to improve the accuracy of the algorithm, we had to include the 'Spam' tag into the list of tags used at the second stage. The list of the tags is introduced below (the description of the tags in provided in the introduction section). The size of data considered at the second stage was 1.24 GB.

1. Buying
2. Signing Up
3. Non Ham (Spam)
4. Personal

After training the algorithm, the size of the database which contains the words and its probabilities was 1620 KB (202 messages). The results obtained after the first stage are as following:



For the purposes of easier understanding, the graph below shows percentage values of the category break-down:



Due to the big amount of data, the algorithm accuracy was computer using 50 messages as following:

$Accuracy = \frac{\#MatchedMessages}{50} * 100\%$; the message is considered to be matched if the tag assigned by human is the same as assigned by the Naïve Bayes Classifier.

$$Accuracy = \frac{41}{50} * 100\% = 82\%$$

5. CONCLUSION

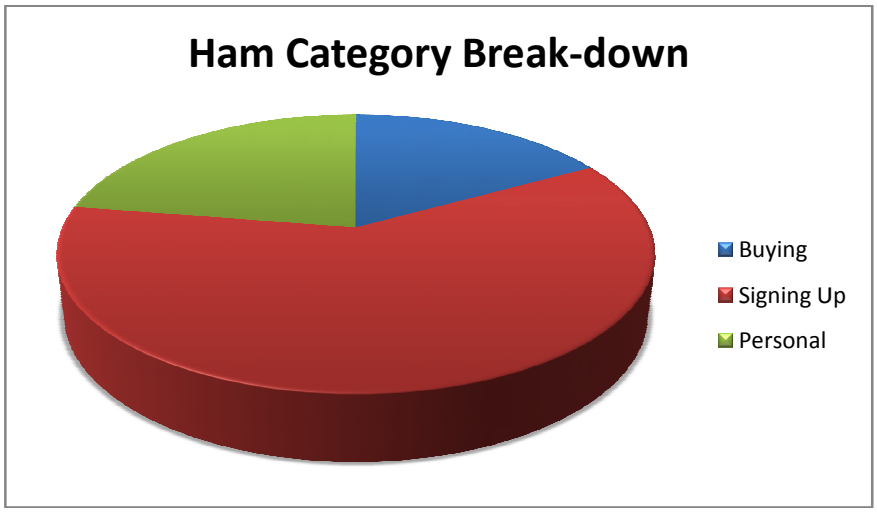
After completing all the stages of the algorithm it's obvious that mostly the DEA services are used for 'spam', which is more than 70%. The minority of usage belongs to the 'ham' category; it's only 6.7% of all the data. The other small usage goes under the category 'other'; it's only 6.2%. It's important to note that a little less 1/5 of the data is used by foreigners (i.e. in language other than English). The detailed breakdown is provided below:

Category	**# of messages**	**% of messages**
Non English	83726	15.13%
Ham	36746	6.66%
Other	33559	6.07%
Spam	387535	70.1%
Errors	11418	2.04%

The main purpose of the paper was to classify and breakdown the 'ham' category of emails, but it couldn't be done without completing two stages of the algorithm and split out the 'spam' data at the second stage. Mostly the 'ham' emails are used for signing up for different services and the percentage of this data is over 60% (4.11% of the whole data set). The second largest subcategory which is being used in the 'ham' category is 'personal'; its percentage is 22.5% (1.53% of the whole data set). The least used subcategory of the 'ham' category belongs to 'buying'; its percentage is roughly around 17% (1.14% of the whole data set). Please refer to the table below for the breakdown of the 'ham' category.

Category	**# of messages**	**% of messages**
Buying	3163	16.9%
Signing Up	11345	60.62%
Personal	4207	22.48%

Please refer to the graphical illustration of the ham category below:



The future work mainly consists of considering the messages that are considered to be in the 'ham' category. A closer look at those messages will provide an image of what are the main purposes and premises of preferring using the DEA services over using standard email accounts for legitimate purposes. Answering the question "Can human readable filter be found among this data to look at" will be answered by manual processing of the split data.

REFERENCES

[1] Smirnov Arthur Describing how the data obtained from DEA services is being used nowadays// Economics and Environmental Management. – 2014

[2] Seigneur, J-M., and Christian Damsgaard Jensen. "Privacy recovery with disposable email addresses." Security & Privacy, IEEE 1.6 (2003): 35-39

[3] Reed, Micheal G., Paul F. Syverson, and David M. Goldschlag. "Anonymous connections and onion routing." Selected Areas in Communications, IEEE Journal on 16.4 (1998): 482-494

[4] Bradley, David (2009-05-13). "Spam or Ham?" Sciencetext. 2011-09-28

[5] Shields, Clay, and Brian Neil Levine. "A protocol for anonymous communication over the Internet." Proceedings of the 7th ACM conference on Computer and communications security. ACM, 2000.

[6] Smirnov, Arthur. "Artificial Intelligence: Concepts and Applicable Uses." Lambert Academic Publishing (2013).

[7] Schneider, Karl-Michael. "A comparison of event models for Naive Bayes anti-spam e-mail filtering." Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003.